

# Об оптимальности методов построения решающих функций

Неделько В. М.

Институт математики СО РАН, г. Новосибирск  
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».  
Лекция 11.

# Аннотация

Рассматривается проблема оценивания качества методов построения решающих функций в задачах анализа данных.

Обсуждается подход, основанный на задании эталонного набора тестовых задач.

# Основные понятия

Пусть  $X$  – пространство значений переменных,  
используемых для прогноза,  
 $Y = \{0, 1\}$  – пространство значений прогнозируемых  
переменных,  
 $\mathcal{C}$  – множество всех вероятностных мер на заданной  
 $\sigma$ -алгебре подмножеств множества  $D = X \times Y$ .

При каждом  $c \in \mathcal{C}$  имеем вероятностное пространство:  
 $\langle D, \mathcal{B}, P_c \rangle$ , где  $\mathcal{B}$  –  $\sigma$ -алгебра,  $P_c$  – вероятностная мера.  
Параметр  $c$  будем называть *стратегией природы*.

# Риск

Решающей функцией (алгоритмом классификации) называется соответствие  $\lambda: X \rightarrow Y$ .

Качество принятого решения оценивается заданной функцией потерь  $\mathcal{L}: Y^2 \rightarrow [0, \infty)$ .

Положим  $\mathcal{L}(y, y') = \begin{cases} 0, & y=y' \\ 1, & y \neq y' \end{cases}$ .

Под риском будем понимать средние потери:

$$R(c, \lambda) = \mathbb{E} \mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x)) \, \mathbb{P}_c(dx, dy),$$

$x \in X, y \in Y$ .

## Функция потерь для оценки вероятности

Возможна более общая постановка задачи, когда под решающей функцией понимается оценка  $\tilde{g}(x)$  условной вероятности

$$g(x) = P_c(y = 1 \mid x) = \frac{P_c(dx, y = 1)}{P_c(dx)}.$$

Качество решения  $\tilde{g}(x)$  можно выражать следующей функцией потерь

$$\mathcal{L}(y, \tilde{g}(x)) = -I(y = 1) \cdot \ln \tilde{g}(x) - I(y = -1) \cdot \ln(1 - \tilde{g}(x))$$

Выборочное среднее данной функции потерь есть взятая со знаком минус функция правдоподобия выборки по отношению к оценке условной вероятности.

# Кривая ошибок

Другим распространённым критерием качества оценки  $\tilde{g}(x)$  является AUC — area under the curve, т.е. площадь под так называемой ROC-кривой (receiver operating characteristic, или кривая ошибок).

Пусть  $F_{\tilde{g}}^y(z)$  — условная функция распределения случайной величины  $\tilde{g}(x)$ , определяемая условной мерой  $P_c(E \mid y)$ , где  $E \subseteq X$  — событие.

Тогда ROC-кривая определяется как кривая, заданная параметрически множеством точек  $(F_{\tilde{g}}^{-1}(z), F_{\tilde{g}}^1(z))$ , когда  $z$  изменяется от  $-\infty$  до  $+\infty$ , и отрезков, соединяющих последовательные точки в случае разрывов функций распределения.

## Свойства кривой ошибок

Начало ROC-кривой в точке  $(0, 0)$ , конец — в  $(1, 1)$ .

Чем больше значение AUC, тем лучше решение  $\tilde{g}(x)$ .

Значение 0,5 соответствует наихудшему качеству  $\tilde{g}(x)$ , учитывая что  $AUC(1 - \tilde{g}(x)) = 1 - AUC(\tilde{g}(x))$ .

Строго монотонное преобразование функции  $\tilde{g}(x)$  не меняет AUC.

Для «обычной» решающей функции значение AUC есть просто среднее арифметическое между вероятностями правильного прогнозирования каждого класса.

На выборке используются эмпирические функции распределения.

# Метод построения решающих функций

Пусть  $Q: D^N \rightarrow \Lambda$  — метод (алгоритм) построения решающих функций,  $\lambda_{Q,V}$  — функция, построенная по выборке  $V$  методом  $Q$ ,  $\Lambda$  — заданный класс решающих функций.

Метод  $\tilde{Q}$ , минимизирующий эмпирический риск, есть

$$\lambda_{\tilde{Q},V} = \arg \min_{\lambda \in \Lambda} \tilde{R}(V, \lambda).$$

# Сравнение методов построения решающих функций

- Выбор эталонного набора тестовых задач.
- Введение понятия оптимальности метода.

## Сопоставление с задачей проверки гипотез

Статистический критерий можно считать частным случаем метода построения решающих функций, когда в роли решающей функции выступает предикат.

В роли риска ошибка второго рода, но функция потерь зависит от распределения.

Известно множество статистических критериев, но нет понятия наилучшего критерия (кроме случая известной простой альтернативы).

## Система «Полигон» — 1980-е

Лбов Г.С., Старцева Н.Г. Сравнение алгоритмов  
распознавания с помощью программной системы «Полигон»  
// Анализ данных и знаний в экспертных системах.  
Новосибирск, 1990. Вып. 134: Вычислительные системы. С.  
56–66.

Принципы:

- для каждого метода включается «эталонная» задача,
- на «своей» задаче метод должен работать лучше других,
- возможно оценить степень универсальности метода,
- тестовая единица - таблица данных.

## Система «Полигон» — 2000-е

Воронцов К.В., Ивахненко А.А., Инякин А.С., Лисица А.В., Минаев П.Ю. «Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации // Всеросс. конф. Математические методы распознавания образов-14 - М.: МАКС Пресс, 2009. С. 503–506.

Принципы:

- использование реальных задач,
- большое число характеристик качества,
- основной критерий - скользящий экзамен.

# Тестовые единицы

Возможные тестовые единицы:

- таблица данных,
- распределение,
- класс распределений.

# Проблема определения оптимальности метода

Напомним, что метод — это отображение выборок в решения.

- Для таблицы данных понятие оптимального метода не имеет смысла.
- Для заданного распределения оптимальный метод вырожден — он любой выборке сопоставляет байесовское решающее правило.
- Об оптимальности метода можно говорить только для класса распределений.
- Даже для нормальных распределений оптимальный метод неизвестен.

# Минимаксный подход к оцениванию качества

- Максимальное по классу распределений значение риска для всех методов одинаково.
- Вводить ограничения сверху на Байесовский уровень ошибки не имеет смысла.
- Использование максимума не абсолютного риска, а отнесённого к достижимому уровню, позволяет ввести осмысленное понятие метода, оптимального на классе распределений.

## Достижимый уровень качества

- Интересует не Байесовский уровень ошибки, а тот, который реально достижим.
- Необходимо задавать для каждого распределения из класса.
- Определяется на основе эталонного метода.

## Выбор классов распределений

- Класс распределений не должен быть ни узким ни широким — иначе получаем соответственно вырожденный метод или аналог NFL.
- Представительность для исследований: все значения достижимого риска, максимально смещённые оценки.
- Параметр сложности. Универсальность.
- Замкнутость относительно допустимых преобразований пространства переменных.

# Варианты классов распределений

- Класс кусочно–постоянных распределений.
- Класс нормальных распределений.
- Класс, сформированный случайными решающими деревьями.
- Ядерные функции для условных вероятностей.

# Выводы

Рассмотрена проблема создания полигона тестовых задач для исследования качества методов построения решающих функций.

Показана целесообразность в качестве тестовых единиц в таком полигоне использовать специальным образом подобранные классы распределений. Распределения подбираются таким образом, чтобы статистическое моделирование на них по возможности полно отражало особенности тестируемого метода обучения.

Это, в частности, означает, что класс распределений должен являться параметрическим семейством, один из параметров которого есть наименьшее достигаемое значение риска в заданном классе решающих правил.

Другим важным параметром является величина смещения эмпирического риска.